Population Genomics Analysis of Legume Host Preference for Specific Rhizobial Genotypes in the *Rhizobium leguminosarum* bv. *viciae* Symbioses

Beatriz Jorrin and Juan Imperial

***Rhizobium leguminosarum* bv. *viciae* establishes root nodule symbioses with several legume genera. Although most isolates are equally effective in establishing symbioses with all host genera, previous evidence suggests that hosts select specific rhizobial genotypes among those present in the soil. We have used population genomics to further investigate this observation. *Pisum sativum*, *Lens culinaris*, *Vicia sativa*, and *V. faba* plants were used to trap rhizobia from a well-characterized soil, and pooled genomic DNA from 100 isolates from each plant were sequenced. Sequence reads were aligned to the *R. leguminosarum* bv. *viciae* 3841 reference genome. High overall conservation of sequences was observed in all subpopulations, although several multigenic regions were absent from the soil population. A large fraction (16 to 22%) of sequence reads could not be recruited to the reference genome, suggesting that they represent sequences specific to that particular soil population. Although highly conserved, the 16S to 23S ribosomal RNA gene region presented single nucleotide polymorphisms (SNP) regarding the reference genome, but no striking differences could be found among plant-selected subpopulations. Plant-specific SNP patterns were, however, clearly observed within the *nod* gene cluster, supporting the existence of a plant preference for specific rhizobial genotypes. This was also shown after genome-wide analysis of SNP patterns.**

Many legumes are able to establish nitrogen-fixing root endosymbiosis with members of a number of proteobacterial groups collectively known as the rhizobia (Vanrhijn and Vanderleyden 1995). The importance of this symbiosis for agriculture and for the nitrogen cycle in the biosphere is widely recognized (Graham and Vance 2003; Herridge et al. 2008). Rhizobia fix nitrogen within a specific organ, the legume root nodule, which represents the outcome of a highly complex developmental program in which both plant and bacteria exchange molecular signals encoded by nodulation and symbiosis-specific genes (Denarie et al. 1996; Gage 2004; Hirsch 1992; Long 1996; Oldroyd et al. 2011). As a result of this molecular signal exchange, legume-rhizobial symbioses are highly specific and only certain rhizobia can establish effective diazotrophic symbioses with any given legume (Denarie et al. 1996; Long 1996; Oldroyd et al. 2011). There are exceptions to this high specificity in the legume-rhizobial symbioses, such as in the case of some tropical legumes that exhibit promiscuous nodulation by a large number of different rhizobia (Martinez-Romero 2003; Perret et al. 2000) or in the case of certain rhizobia that are equipped with different sets of nodulation genes specific for different plants (Koch et al. 2010; Perret et al. 2000).

A different situation arises when a given rhizobium equipped with just one set of symbiotic and specificity determinants can establish effective symbioses with different legume hosts. This is the case of *Rhizobium leguminosarum* bv. *viciae*. This bacterium is able to establish effective symbioses with members of the *Pisum*, *Vicia*, *Lens*, and *Lathyrus* genera, all within the tribe Vicieae (Doyle and Luckow 2003), and contains one set of nodulation and nitrogen fixation genes harbored in a plasmid responsible for the symbiotic genotype (Sym plasmid) (Surin and Downie 1989; Young et al. 2006). With some exceptions (Mutch et al. 2003), most *R. leguminosarum* bv. *viciae* isolates containing a Sym plasmid are able to efficiently nodulate members of all four plant genera named above. However, it has been suggested that, within *R. leguminosarum* bv. *viciae*, some genotypes exist that are better adapted to specific plants and that the plant prefers, and thus selects for, these genotypes among those available in a given soil. Evidence in favor of this plant selection hypothesis has been obtained in the past through the use of restriction fragment length polymorphism analysis of polymerase chain reaction–amplified (PCR-RFLP) symbiotic and nonsymbiotic molecular markers from isolates obtained from nodules of different plant genera (Depret et al. 2004; Laguerre et al. 2003; Louvrier et al. 1996; Mutch and Young 2004; Palmer and Young 2000). These studies relied on just a few, arbitrarily selected markers and provided little information on the nature of the genotypes selected by specific plants or on the basis for their selection. We have now taken advantage of the availability of both cost-effective next generation sequencing technologies and a reference *R. leguminosarum* bv. *viciae* genomic sequence (Young et al. 2006) to reappraise the possible selection of specific *R. leguminosarum* bv. *viciae* genotypes by their different legume plant hosts. Toward this aim, we have used a population genomics approach (Futschik and Schloetterer 2010; Kofler et al. 2011a and b; Schloetterer et al. 2014) that we have adapted to

the study of rhizobial endosymbiont populations (Jorrin and Imperial 2014). We use this methodology to genomically compare rhizobial populations selected by *Pisum sativum*, *Lens culinaris*, *Vicia sativa*, and *V. faba* plants from a well-characterized soil and provide genomic evidence for plant-determined enrichment of specific *R. leguminosarum* bv. *viciae* populations.

Table 1. Average coverage in RPKM of each reference genome (*Rhizobium leguminosarum* bv. *viciae* 3841) replicon by Pool-Seq DNA samples of the four plant-selected populations

Replicon	Pea	Lentil	Fava bean	Vetch
Chromosome	121.0	112.9	108.8	120.3
pRL7	8.3	7.6	6.8	6.3
pRL8	7.5	6.2	6.7	6.0
pRL9	95.2	84.6	96.7	84.2
pRL10	90.1	74.1	86.5	74.8
pRL11	90.0	79.3	84.9	80.7
pRL12	109.3	97.4	98.4	99.9

[a] RPKM = reads per kilobase per million reads.

RESULTS

Plant-selected populations.

Plant-selected *R. leguminosarum* bv. *viciae* populations consisted of 100 isolates obtained, one each, from surface-sterilized root nodules excised from trap plants inoculated with our experimental P1 soil. Isolates were plate-purified, were retested for symbiotic properties, and were stored. Four populations were obtained, i.e., pea (*P. sativum*), lentil (*L. culinaris*), vetch (*V. sativa*), and fava bean (*V. faba*).

Next generation sequencing of pooled, plant-selected rhizobial populations.

DNA from pooled samples of each of the four different plant-selected *R. leguminosarum* bv. *viciae* populations were submitted to next-generation sequencing (Pool-Seq) so that, on average, each of the 100 genomes was sequenced to 2 to 3× coverage (Illumina HiSeq2000, 180 bp PE libraries, 100 bp reads, 12 Mreads; BGI Hong Kong). This figure was chosen as a compromise between the higher sequencing costs at the time

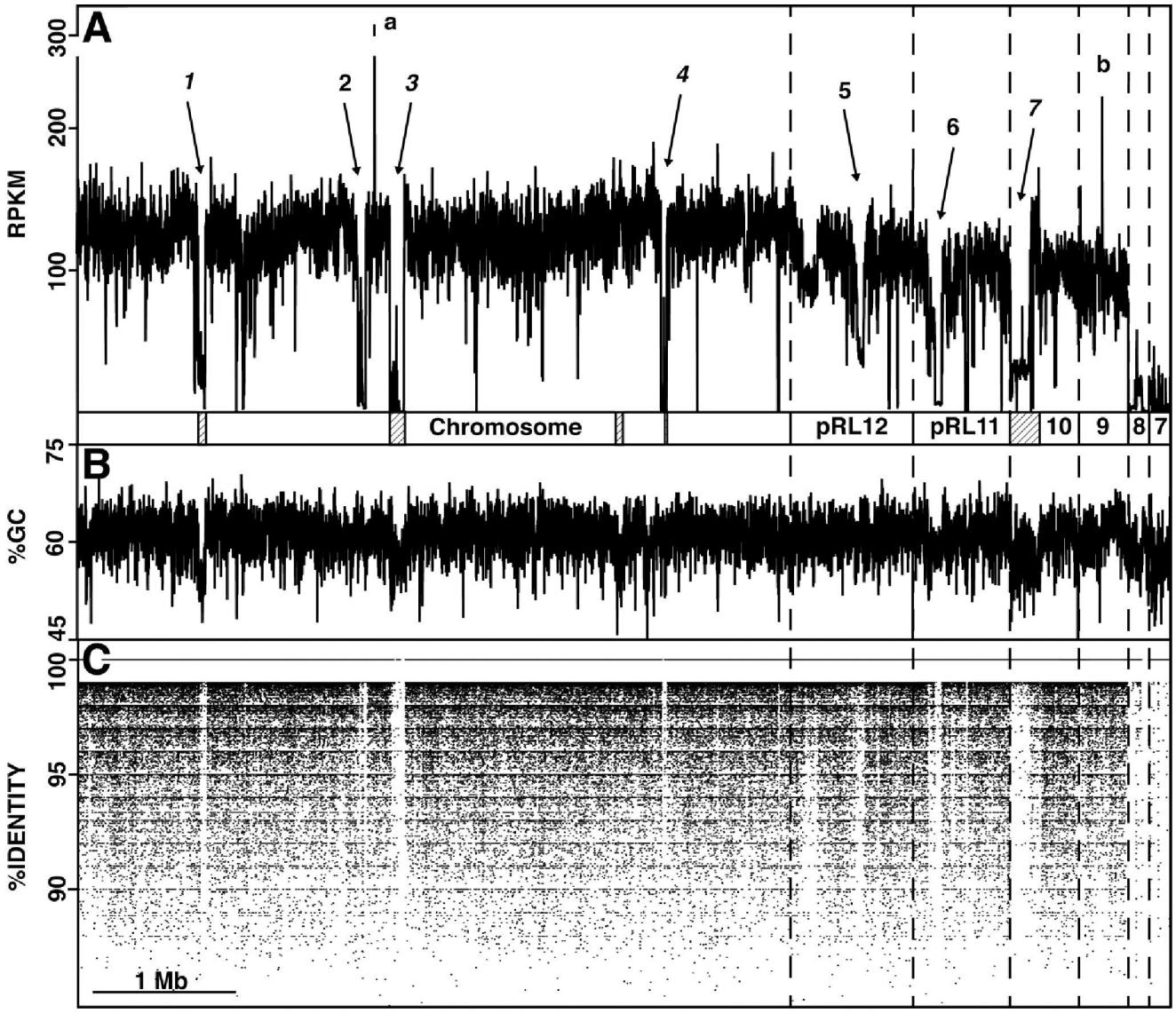

Fig. 1. Conservation of *R. leguminosarum* bv. *viciae* 3841 reference genome sequences in the P1 soil rhizobial population. **A,** Coverage (reads per kilobase per million reads) of genome regions by Bowtie2 recruited reads. The location and span of five low G+C islands described by Young and associates (2006) in the reference genome are indicated by hatched bars. **B,** Percent G+C plot of genomic regions. **C,** Identity (%) of recruited reads. Sequences from all the reference genome replicons were concatenated as shown. Major underrepresented and overrepresented regions in the P1 soil dataset are indicated by numbers and lowercase letters, respectively.

(2011) and the ability to detect sequences that might be present in just one of the isolates forming the population and resulted in an average coverage of 200 to 240× for sequences present in the reference *R. leguminosarum* bv. *viciae* 3841 reference genome (Jorrin and Imperial 2014; data not shown). Individual sequencing reads from each population were recruited against the reference genome (Jorrin and Imperial 2014), and coverage results were normalized to correct for minor differences in sequencing depth among the four Pool-Seq datasets.

Presence of sequences from the *R. leguminosarum* bv. *viciae* reference genome in Pool-Seq datasets.

The *R. leguminosarum* bv. *viciae* 3841 reference genome is multipartite and contains six large plasmids: pRL7 (151.5 kb) through pRL12 (870 kb), in addition to a 5,057-kb circular chromosome (Young et al. 2006). The presence and conservation of sequences from these seven replicons in the Pool-Seq datasets from P1 soil was investigated (Table 1). Sequences from the two smaller plasmids, pRL7 and pRL8, were minor components of the Pool-Seq datasets, suggesting that these plasmids, if present, are a minor component of the P1 soil *R. leguminosarum* bv. *viciae* population. The highest coverage was found with the chromosome (108.8 to 121.0 reads per kilobase per million reads [RPKM]), followed by plasmids pRL12 (97.4 to 109.3 RPKM) and pRL9 (84.2 to 96.7 RPKM). The highest coverage values for the four largest plasmids (pRL9 to pRL12) were found in the pea dataset, and the lowest in the lentil and vetch datasets. These data suggest that most of the sequences in the chromosome and in the larger plasmids of the reference strain were present and

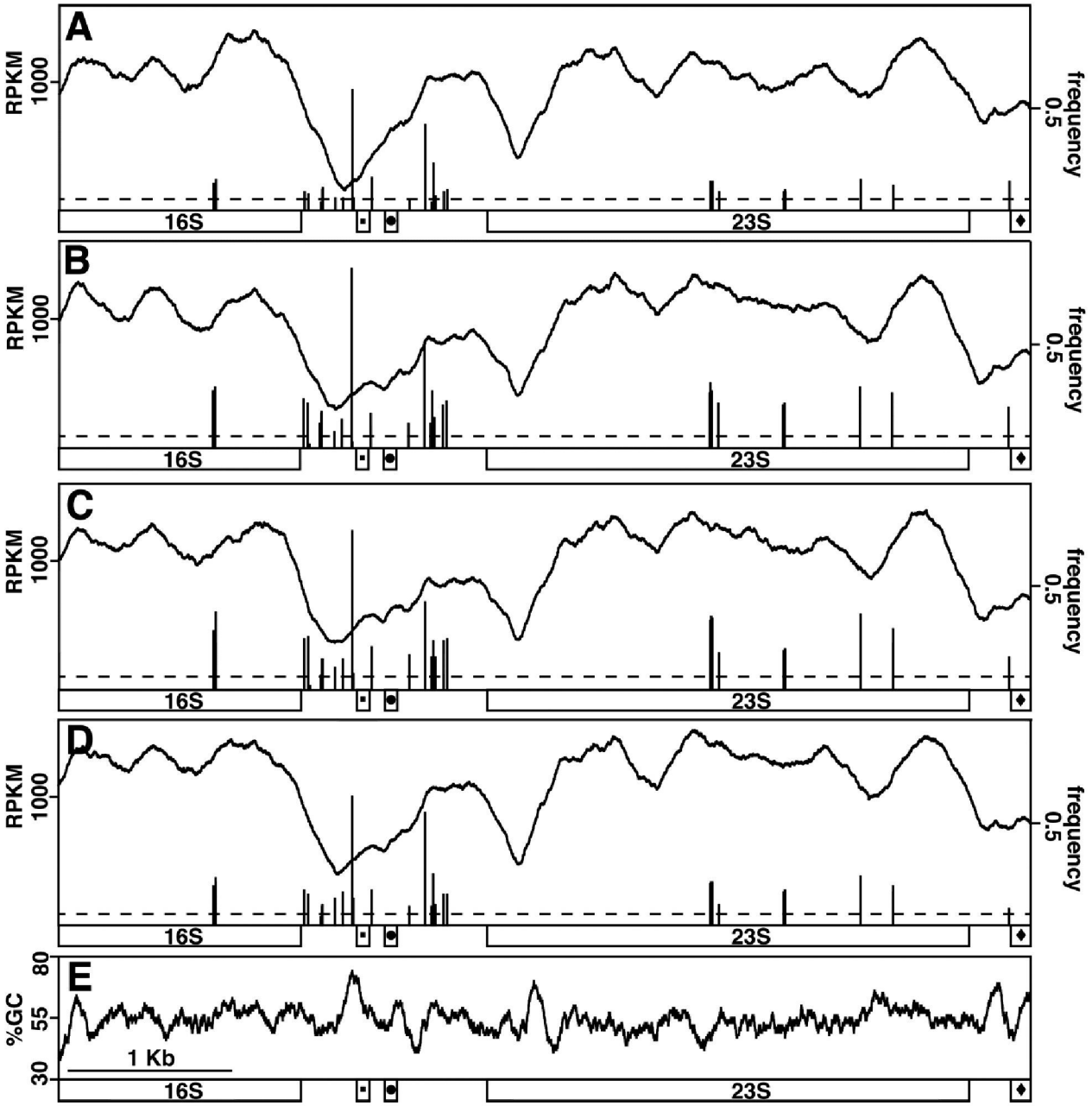

Fig. 2. Conservation and diversity of the 16S-23S rRNA genomic region in the four plant-slected rhizobial populations: **A,** pea; **B,** lentil; **C,** fava bean; and **D,** vetch. Coverage (reads per kilobase per million reads) along the reference genome sequence is indicated by a continuous line. Single nucleotide polymorphism location and frequency are indicated by vertical lines. A discontinuous horizontal line indicates the $P < 0.01$ threshold, which, in these experiments, corresponded to a coverage of approximately 5%. The position of Ile-tRNA, Ala-tRNA, and 5S rRNA in the reference genome sequence is indicated by boxes containing a small square, a large dot, and a diamond, respectively. **E,** Percent G+C composition along the genomic region.

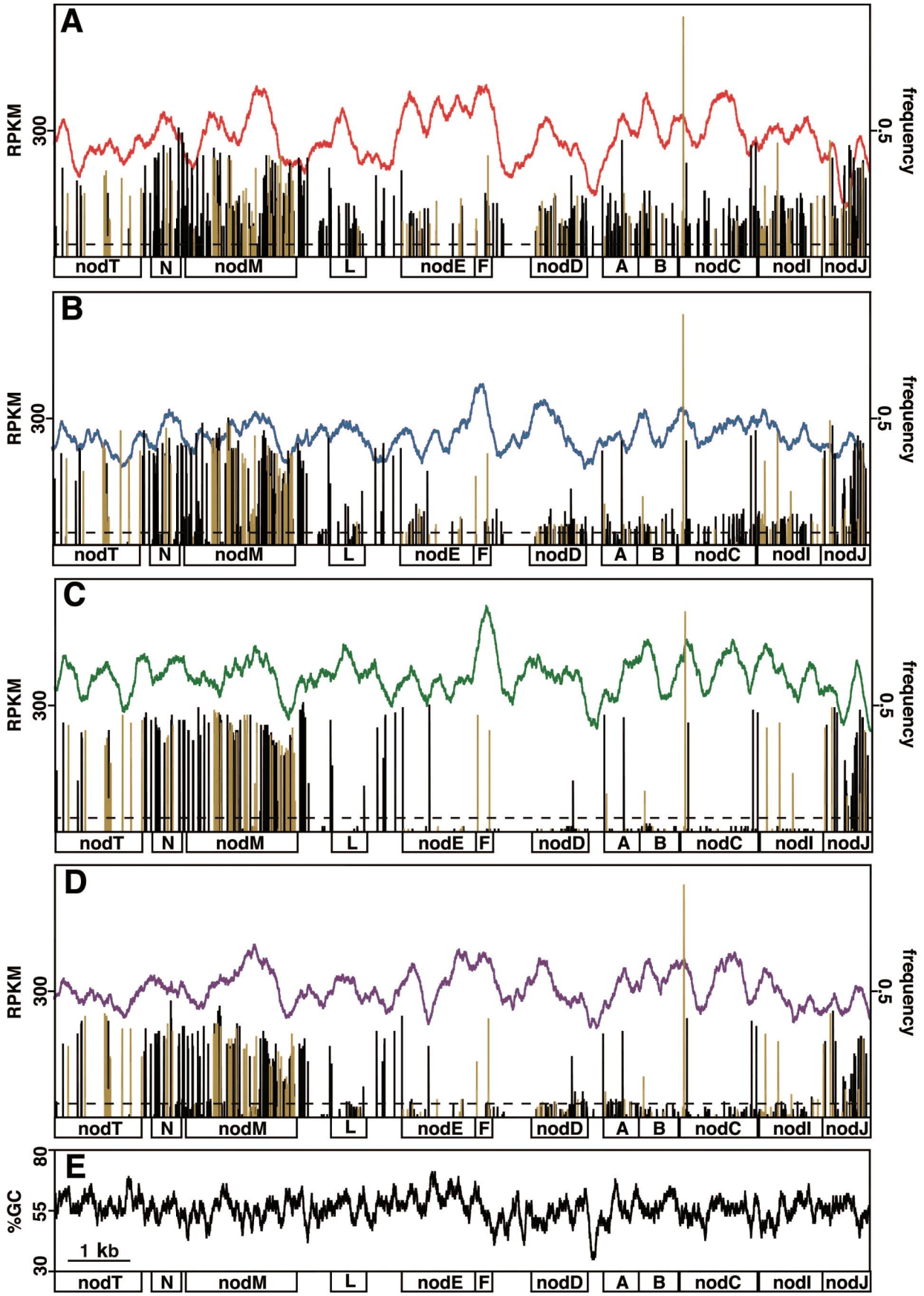

Fig. 3. Conservation and diversity of the *nod* genomic cluster in the four plant-selected rhizobial populations: **A,** pea; **B,** lentil; **C,** fava bean; and **D,** vetch. Coverage along the reference genome sequence is indicated by a continuous line. Single nucleotide polymorphism location and frequency are indicated by black (synonymous substitutions) or golden (nonsynonymous substitutions) lines. A discontinuous horizontal line indicates the $P < 0.01$ threshold, which, in these experiments, corresponded to a coverage of approximately 5%. **E,** Percent G+C composition along the genomic region.

well conserved in the P1 soil population but, also, that differences in coverage exist among plant-selected populations. These differences could be explained in terms of differential divergence or absence of specific genes or regions from the reference genome in the genomes of members of the different plant-selected populations and were analyzed by examining the coverage of the reference genome replicons by recruited sequence reads of each of the four Pool-Seq datasets. Plots of coverage (RPKM) along replicon sequences are presented in Supplementary Figures S1 to S6 for the chromosome, plasmids pRL7 and pRL8, pRL9, pRL10, pRL11, and pRL12, respectively. No striking differences in coverage among the four Pool-Seq datasets were observed for any of the replicons, although some regions from the reference genome appeared to be absent or overrepresented in the P1 soil populations. These regions are shown in Figure 1, in which sequence reads from all four datasets were pooled and recruited against the reference genome replicons and coverage and read sequence identity were plotted. Only two peaks of overrepresented contiguous sequences (coverage above 200 RPKM) appeared. These corresponded to chromosomal gene RL2002 (a DDE transposase, peak a) and to gene pRL90159 (a hypothetical protein, peak b) present in pRL9. Conversely, seven regions from the reference genome were absent or near absent in the P1 soil datasets. In addition, plasmids pRL8 and pRL7 were nearly absent, with very few reads being recruited to them. Regions 1 (RL0791 to RL0841), 2 (RL1869 to RL1944), 3 (RL2105 to RL2195), and 4 (RL3912 to RL3956) were low-coverage chromosomal regions, spanning 51, 76, 91, and 45 genes, respectively. These were genes of differing annotated functions, including enzymes, transmembrane or transport proteins, transcriptional factors, hypothetical proteins, and other functions including insertion sequences (region 1), transposases (regions 2 and 3), and phage proteins (region 4). The remaining three low-coverage regions were located, one each, in the three larger plasmids. Region 5 spanned 40 genes (pRL120440 to pRL120479) and included a 14-gene *imp* cluster (*impA* through *impN*) that contains a type 6 secretion system. Region 6 contains 44 genes (pRL110145 to pRL110188) of different annotated functions. Region 7 is present in the symbiotic plasmid pRL10 and spans 128 genes (pRL100011 to pRL100138) flanked by an integrase and a recombinase, suggesting that it could represent mobile DNA. In addition to genes annotated as encoding different enzyme activities and phage proteins, transposases and a recombinase/integrase, transcriptional regulators, a cold-shock protein, and a *repA* gene are present in region 7. All seven low-coverage regions shown in Figure 1A correspond to regions recruiting few, low-identity reads from the datasets (Fig. 1C). However, with the exception of part of regions 3 and 4, all seven regions recruit some reads at 100% identity (Fig. 1C), suggesting that these genomic regions are present in a minor fraction of the *R. leguminosarum* bv. *viciae* P1 soil population. The same situation was found for plasmid pRL7 and part of plasmid pRL8 (Fig. 1C). Underrepresented regions often corresponded to low G+C regions (Fig. 1B). This was clearly the case for regions 1, 3, 6, and 7. In addition, four of these regions (1, 3, 4, and 7) correspond to four of the five low G+C genomic islands identified by GC3 analysis by Young and associates (2006).

Despite the large conservation of the reference genome within the P1 soil population, not all sequence reads from the subpopulation Pool-Seqs were recruited by the reference genome, and in fact, the unrecruited sequences unique to the P1 soil datasets constitute a large fraction of the reads (15.8, 22.2, 18.1, and 22.8% for the pea, lentil, vetch, and fava bean subpopulations, respectively). Aside from possible artefactual sequences, these sequence reads probably arise from genomic regions present in the P1 soil population but absent in the reference genome.

Sequence polymorphisms in specific regions of Pool-Seq datasets.

Previous work on plant host selection of *R. leguminosarum* bv. *viciae* genotypes had centered on PCR-RFLP analyses of the *nodD* to *nodF* region (Mutch and Young 2004; Laguerre et al. 2003). The Pool-Seq methodology allows the comparative study of population-linked polymorphisms at the single nucleotide level (single nucleotide polymorphisms [SNP]) for any gene or region of interest in the reference genome. We analyzed two contrasting regions, the chromosomal 16S to 23S ribosomal (r)RNA region encoding conserved housekeeping rRNAs and the symbiotic plasmid-encoded *nod* cluster responsible for synthesis of the Nod factor, the main symbiotic signal.

16S to 23S rRNA region.

The reference *R. leguminosarum* bv. *viciae* 3841 genome contains three copies of the rRNA operon, all identical and located in the chromosome, with the structure: rRNA 16S-tRNA Ile-tRNA Ala-rRNA 23S-rRNA 5S (Young et al. 2006). Plant-selected Pool-Seq datasets were recruited to this region, and coverage and SNP were plotted along the region (Fig. 2). The high level of coverage for all four plant-selected datasets, above 1,000 for 16S and 23S rRNAs, suggests that most rhizobia in these populations contain several copies of these genes, perhaps three, as in the reference strain or in other *R. leguminosarum* strains (*R. leguminosarum* bv. *trifolii* WSM2304, accession number PRJNA20179; *R. leguminosarum* bv. *trifolii* WSM1325, PRJNA20097; *R. leguminosarum* bv. *viciae* TOM, PRJNA199010; *R. leguminosarum* bv. *trifolii* SRDI943, PRJNA199021; *R. leguminosarum* bv. *phaseoli* 4292, PRJNA199148; *R. leguminosarum* bv. *trifolii* SRDI565, PRJNA199011; *R. leguminosarum* bv. *viciae* 248, PRJNA201173; *R. leguminosarum* bv. *viciae* WSM1481, PRJNA199020; *R. leguminosarum* bv. *trifolii* CB782, PRJNA67103; *R. leguminosarum* bv. *trifolii* WSM1689, PRJNA62289; *R. leguminosarum* bv. *trifolii* WSM2012, PRJNA18209). Coverage of the reference sequence decreased drastically for all four plant-selected populations in the 16S to 23S intergenic region, suggesting that parts of this region may be nonconserved or absent in rhizobia from the P1 soil. This was also true of the Ile- and Ala-tRNAs present in the 16S to 23S intergenic region of the reference strain, especially for the Ile-tRNA, which appeared to be nearly absent in the pea Pool-Seq dataset (Fig. 2). Comparative genomic analysis of rRNA clusters from other sequenced *R. leguminosarum* strains (discussed above) indicated a generalized lack of conservation in the 16S to 23S region, although Ile- and Ala-tRNAs were present in most cases (data not shown). It is worth noting that decreases in sequence coverage also occurred in two intragenic regions of the 23S rRNA for all four Pool-Seq populations, although we do not currently have a clear explanation for this observation. One possibility is that the Illumina sequence technology could be affected by sequence composition, and Aird and associates (2011) showed that the PCR step in sequencing library preparation selects strongly against regions with anomalous percent G+C. Indeed, inspection of the percent G+C plot (Fig. 2E) suggested that regions of low coverage are associated with high percent G+C peaks. Alternating high- and low-percent G+C regions were also common in 23S rRNA genes from other *R. leguminosarum* genomes (data not shown). Despite these anomalies, the overall coverage pattern for 16S and 23S rRNA sequences was very similar for all four plant-selected datasets.

As predicted, few SNP were recorded in this region (Fig. 2). Pairwise fixation indices (F_{ST})were low ($F_{ST} < 0.1$) with $P <$

0.001 for all SNP except for a single SNP in the pea/lentil comparison (data not shown), indicating that observed SNP were not specific for any of the plant-selected populations. In all, 32 polymorphic sites along a 5,903-bp stretch (0.54%) were observed, two in the 16S rRNA gene, 11 in the 23S rRNA gene, and 19 in the 16S to 23S intergenic region. Except for a specific SNP in the 16S to 23S intergenic region, no SNP was present at frequencies higher than 0.5, suggesting that the reference genome sequence represents the most abundant genotype for this region in the P1 soil rhizobial population. Some differences were observed in the relative frequency of these SNP in the four populations, with lower frequencies in the pea population, which may be interpreted as suggesting a preference for the reference genome genotype, which was, in fact, isolated from a pea nodule (Young et al. 2006).

nod cluster.

A similar analysis of Pool-Seq populations was carried out with the *nod* region. Figure 3 shows the read coverage and SNP distribution along a stretch of 13,462 bp from symbiotic plasmid pRL10 that includes the *nodTNMLEFDABCIJ* genes. The coverage pattern was similar among the four plant-selected populations, but important differences were observed in the number and distribution of SNP. A total of 455 polymorphic sites were recorded (3.4%), with 409 in coding regions and 46 in intergenic regions. Overall, a large concentration of polymorphic sites was observed within the *nodTMN* and *nodJ* coding sequences. This included both synonymous and nonsynonymous sites, with polymorphism frequencies of up to 0.5 with respect to the reference genome but with little difference among Pool-Seq datasets. Clear differences, however, were observed among plant-selected populations in the central *nod* cluster *nodLEFDABCI*, especially between the pea and fava bean datasets. Along this region, the fava bean dataset showed a small number of polymorphic sites and at low frequencies, except for a few that appeared in all four datasets and that probably characterize the P1 soil rhizobial population. In this respect, it is worth mentioning that a nonsynonymous (Ala > Ser) substitution (G > T) at position 55 of the *nodC* coding region was present in all four populations at frequencies of approximately 0.9, and this could be taken as characteristic of the P1 soil *nod* cluster (although this substitution is present in many of the *R. leguminosarum nodC* sequences in GenBank; data not shown). In contrast, the pea dataset presented a large number of low-frequency (0.2 to 0.3) SNP within the central *nod* cluster. A similar pattern was observed for the lentil population, although the SNP frequencies were lower, whereas the vetch population displayed a pattern intermediate between those of lentil and fava bean. Significance analysis of the observed SNP among plant-selected subpopulations was carried out using their pairwise F_{ST} indices (Supplementary Table S1). Whereas no SNP with $F_{ST} > 0.1$ was observed for the lentil/fava, lentil/vetch, or fava/vetch pairs, and only one for the pea/lentil pair, high (85) or intermediate (8) numbers were observed for the pea/fava and the pea/vetch pairs, respectively, thus confirming the genotypic differentiation among these subpopulations. As expected, all P1 soil-derived populations showed large differences with the reference strain 3841 (151 to 192 SNP with $F_{ST} > 0.1$ and $P < 0.002$).

Genome-wide sequence polymorphisms in Pool-Seq datasets.

A comparative analysis of all SNP in sequences recruited by the reference genome was carried out and the data are summarized in Table 2. Similar total numbers of polymorphic sites were observed for all four Pool-Seq P1 soil subpopulations, with the pea subpopulation exhibiting the lowest number (214,467) and the fava bean subpopulation the highest (299,099). These numbers correspond to 2.8 to 3.9% polymorphic sites and can probably be taken as characteristic of the P1 soil population. The levels of polymorphism are comparable albeit somewhat lower than those found within *Rhizobium etli* (4 to 6%; Acosta et al. 2011). For each replicon, they showed slight variations among subpopulations, with pRL7 and pRL8 showing the least polymorphisms (0.5 to 1.0%). This is probably related to the very low abundance of these plasmids in the P1 soil population (Fig. 1). Significance analysis of the observed SNP for the whole genome among plant-selected subpopulations was carried out using their pairwise F_{ST} indices (Supplementary Table S2). An overall pattern similar to that already observed with the *nod* region (discussed above) reappeared at the genomic level. Very few significant SNP (1,264 to 1,593) were observed in pairwise combinations not involving the pea subpopulation, which showed a larger level of differentiation with the vetch and lentil subpopulations (2,485 and 3,998 significant SNP, respectively) and, especially, with the fava subpopulation (11,087 significant SNP). Again, all P1 soil-derived subpopulations showed large differences with reference strain 3841 (106,384 to 223,009 significant SNP).

Nucleotide frequencies at all sites in the reference genome for each subpopulation were transformed into Euclidean distances (Supplementary Table S3) and were subjected to multidimensional scaling and the two main coordinates were plotted in a two-dimensional space (Supplementary Fig. S7). The plot separated all four subpopulations among themselves and from the reference genome. The P1 population as a whole was set apart from the reference genome along dimension 1, whereas dimension 2 clearly separated the most divergent subpopulations, pea and fava bean.

DISCUSSION

The fact that symbiotic *Rhizobium leguminosarum* bv. *viciae* strains, irrespective of their plant origin, are able to nodulate effectively all their legume hosts (genera *Pisum*, *Lens*, *Vicia*, and *Lathyrus*) complicates studies aimed at establishing the existence of a preference on the part of the legume host for specific rhizobial genotypes, and previous studies have relied on arbitrarily selected genotypic markers (Depret et al. 2004; Laguerre et al. 2003; Louvrier et al. 1996; Mutch and Young 2004; Palmer and Young 2000). We have adapted to rhizobial populations (Jorrin and Imperial 2014) the Pool-Seq population genomics approach first developed by the Schloetterer lab for the genomic study of *Drosophila* populations (Futschik and Schloetterer 2010; Kofler et al. 2011a and b; Schloetterer et al. 2014), in order to approach this problem. As discussed, this methodology takes advantage of the existence of a *R. leguminosarum* bv. *viciae* reference genome (Young et al. 2006) and of the diminishing costs of next generation sequencing to genomically compare different populations without the need to

Table 2. Genome-wide abundance of single nucleotide polymorphisms in Pool-Seq DNA samples of the four plant-selected populations[a]

Replicon	Pea	Lentil	Fava bean	Vetch
Total	214,647	269,749	299,099	261,282
Total %	2.77	3.48	3.86	3.37
Chromosome	2.76	3.59	3.98	3.45
pRL7	1.00	0.74	0.94	0.72
pRL8	0.60	0.55	0.51	0.50
pRL9	2.98	3.59	4.00	3.58
pRL10	2.96	3.20	3.43	3.10
pRL11	3.12	3.66	4.09	3.68
pRL12	3.00	3.78	4.21	3.66

[a] Expressed as total number and as percentages of each replicon and of the total reference genome sequence.

choose any specific marker, in itself a limitation, or without the labor and expense associated with sequencing the genome of large numbers of isolates from the different populations (Jorrin and Imperial 2014; Schloetterer et al. 2014).

The source of our *R. leguminosarum* bv. *viciae* populations was soil P1, well-characterized for rhizobial populations in previous studies of plant preference for specific genotypes (Laguerre et al. 2003; Louvrier et al. 1996; G. Laguerre personal *communication*). This allowed us to compare the results of our work with those studies. At the population level, the P1 soil symbiotic *R. leguminosarum* bv. *viciae* population showed an overall good genomic conservation with the reference genome, particularly in the chromosome but also in the large plasmids. Despite this conservation, a few over- and underrepresented regions were identified. In the first case, they correspond to specific genes (a DDE transposase and a hypothetical protein) that probably underwent duplication within the P1 population. In the second case, seven major regions underrepresented in P1 soil correspond to groups of 14 to 127 genes that probably represent acquisitions of the reference genome for ecological adaptations to its habitat. It is noteworthy that four of these regions correspond to low G+C genomic islands previously identified in the reference genome (Young et al. 2006). Among these low G+C islands, the largest (region 7) is located in the symbiotic plasmid (pRL10), just upstream of the *nod* cluster. Among the underrepresented regions not clearly associated to low G+C, it is worth noting the presence of a type 6 secretion system in plasmid pRL12 (region 5). Two smaller plasmids, pRL7 and pRL8, also appeared to be strongly underrepresented in the P1 soil population, although sequences recruited by these plasmids had low levels of polymorphism, suggesting that they are very strongly conserved within the small fraction of cells that carry them.

It is also important to mention that a sizeable number (15.8 to 22.8%) of sequence reads from the Pool-Seq subpopulations were not recruited by the reference genome. Since many of these reads probably represent specific genes present in the P1 soil population and relevant for adaptation to specific ecological conditions in P1 soil, their nature and conservation among subpopulations are currently under study.

The Pool-Seq approach allows detailed analysis of SNP for any specific region of interest within the reference genome. In this study, we chose the highly conserved 16S to 23S rRNA region and the symbiotic nodulation cluster (*nod* genes). Since ribosomal RNAs are subjected to high structural-functional constraints, polymorphisms in these molecules are kept to a minimum and reflect the evolutionary history of the taxons under study (Woese 1987). In fact, rRNA sequence comparisons are at the root of present-day prokaryotic phylogenies (Woese et al. 1990). The 16S and 23S rRNA genes were highly conserved when comparing P1 soil subpopulations and the reference genome, although large differences appeared in the intergenic regions, including the Ile- and Ala-tRNA genes located in the 16S to 23S rRNA intergenic region. Apart from some distinct SNP, a generalized decrease of coverage in this region probably means that it is absent in some of the members of the population. This could be especially true of the Ile-tRNA gene, whose coverage practically disappears in the pea subpopulation. It is possible that the number, position, and frequency of the observed SNP and possible deletions in the intergenic regions may constitute a fingerprint of the soil population, and we are currently testing this in populations from other soils. However, none of the observed differences was specific of any of the plant-selected subpopulations, which suggests that, as expected, rRNA genes are neutral in this respect. The opposite situation was encountered in the analysis of the *nod* gene cluster. These genes are responsible for synthesis of the lipochitoolygosaccharide signal molecule (Nod factor) specifically recognized by the plant (Denarie et al. 1996), and clear differences were observed in the frequency of SNP in this region, both for the P1 population as a whole and for each of the four plant-selected P1 soil subpopulations. Two large polymorphic subregions, *nodTMN* and *nodJ*, differentiated the P1 population from the reference genome, irrespective of the plant-selected subpopulation. However, SNP analysis of the *nodLEFDABCI* region showed clear differences among plant subpopulations, with a large number of SNP in the pea subpopulation, a very small number of SNP in the fava bean subpopulation, and intermediate results for the lentil and vetch subpopulations. It is tempting to interpret these results as meaning that fava bean plants are more restrictive regarding the *nod* genotype of their rhizobial symbionts than pea plants or, alternatively, that the *nod* genotype from the reference genome colonizes fava beans more efficiently, whereas none of the available genotypes shows a clear advantage in the colonization of pea plants. However, any further interpretation is hampered by limitations inherent to the Pool-Seq approach, in which the strain origin of the different reads and, hence, SNP is not known. For instance, for the pea subpopulation, in which many low-frequency SNP loci were observed in the *nod* region, it would be equally possible, in principle, that these SNP are all present within the genomes of a small, highly polymorphic fraction of the population or that they are the sum of polymorphic sites scattered throughout most of the population. At any rate, the observed, contrasting *nod* pattern confirms and extends previous studies with *nodD* and the *nodD* to *nodF* intergenic region in which isolates from fava bean nodules were shown to present lower diversity than isolates from pea nodules (Laguerre et al. 2003; Mutch and Young 2004).

Irrespective of the neutral (such as the 16S to 23S rRNA region) or selected (such as the *nod* region) nature of the specific genomic region considered, we hypothesized that any plant selection of rhizobial genotypes would be reflected in the number or distribution of SNP at the level of the complete genome. Differences in the number and replicon distribution of SNP for the four plant Pool-Seq datasets were observed (Table 2), although their biological significance is not clear.

Although this study and a previous report (Jorrin and Imperial 2014) represent the first Pool-Seq studies of rhizobial populations, other population genomics studies have been undertaken with rhizobia that reveal genomic adaptations to the host and suggest that host selection of rhizobial microsymbiont genotypes could be a general phenomenon in the legume-rhizobium symbiosis. Using low-coverage (0.4 to 1.2×) genome sequencing of 12 *Sinorhizobium medicae* strains isolated from *Medicago lupulina* nodules, Bailly and associates (2011) were able to identify population-specific polymorphisms in conserved genes as well as putative population-specific genes when compared with the *S. medicae* WSM 419 reference strain (isolated from *Medicago murex*). More recently (Epstein et al. 2012; Sugawara et al. 2013), deep sequencing of up to 48 genomes from *S. meliloti*, *S. medicae*, and three other *Sinorhizobium* genospecies was used to study genomic adaptations to the plant host. These studies revealed not only that specific genes, among them *nod* genes and type III, IV, and VI secretion systems, differ among rhizobial genospecies showing different plant specificities but also that clear phenotypic interactions exist among rhizobial genotypes and *Medicago truncatula* host plant genotypes, a clear evidence for rhizobial genotype selection by the plant host.

Our work exemplifies the potential of the PoolSeq population genomics approach to study the selection of specific rhizobial genotypes by the host plant, but its limitations make it difficult to investigate in depth the exact nature of the selec-

tion, in particular the role of any specific genes in this selection. In order to solve this uncertainty, we are currently carrying out genome sequencing of a representative subset of the plant-selected rhizobial isolates. These genome sequences will allow a better characterization of the plant-specific genotype selection, particularly for those genes that are not present in the reference genome.

MATERIALS AND METHODS

Site and soil characteristics.

Soil P1 was collected in the Institute for Agronomy Research Experimental Station of Époisses (Domain expérimental d'Époisses, Bretenière, Côte d'Or, France) and was made available to us by G. Laguerre. Portions of soil were collected with a shovel from several sites within the P1 plot after removing the upper 3 to 5 cm of crust and down to 20 to 30 cm depth. All portions were thoroughly mixed and 100 kg was shipped in a container. P1 soil is a clay loam soil with a pH of 8.0 and its characteristics and resident *R. leguminosarum* bv. *viciae* natural populations have been studied in the past (Louvrier et al. 1996; G. Laguerre *personal communication*). No rhizobial inoculation has been carried out in the past in field P1, but peas were planted in the 1982 season (Laguerre et al. 2003). Afterwards, several nonlegume crops (mainly corn and wheat; G. Laguerre *personal communication*) were planted in field P1 up until sampling (September 2010). A stable population of pea-nodulating *R. leguminosarum* bv. *viciae* (approximately 10^4 to 10^5 rhizobia per gram of soil) has been estimated in soil P1 over time (Louvrier et al. 1996; G. Laguerre *personal communicatio*n).

Trap plant assay.

Pea (*Pisum sativum* L. cv. Frisson), lentil (*Lens culinaris* Medik. cv. Magda), fava bean (*Vicia faba* L. cv. Muchamiel), and vetch (*V. sativa* L. cv. Senda) seedlings were obtained from surface-sterilized seeds (Sanchez-Canizares et al. 2011). Seeds were germinated on agar plates (1%) at 20°C in darkness for 3 days. Trap plants (three plants per Leonard jar, four Leonard jars per trap plant species) were cultivated for about 3 weeks in Leonard jars containing Jensen's solution (Somasegaran 1994). Each jar was filled with a mixture (1:1) of soil and sterile vermiculite. Jars without soil were used as negative nodulation controls.

***R. leguminosarum* bv. *viciae* isolation.**

Endosymbiotic bacteria were isolated from nodules induced in *P. sativum*, *L. culinaris*, *V. faba*, and *V. sativa* trap plants by bacteria present in soil samples. Typical, red nodules were picked from 12 trap plants of each species after 3 weeks and were surface-sterilized and crushed aseptically, and bacteria were isolated and purified on yeast mannitol (YM) agar plates following standard techniques (Vincent 1970). The strains were grown on YM at 28°C and were maintained at –80°C in 20% of glycerol for long-term storage.

PoolSeq sample and DNA extraction.

A total of 100 bacterial isolates were selected from each host. Each of those isolates was independently grown on YM broth for 2 days and each culture was adjusted to an optical density at 600 nm = 2. Portions (1.5 ml) of cultures from isolates from the same host were pooled together and were centrifuged (10 min, 5,000 × *g*). The cell pellet obtained from the combined cell suspension was used to extract genomic DNA (Wilson 2001). Pooled DNA samples were quantitated by Nanodrop (Thermo Scientific) spectrophotometry and Qubit (Life Technologies) fluorometry.

Sequencing and bioinformatic analysis.

The pooled rhizobial DNA from each host was sequenced externally by BGI Hong Kong (Illumina Hi-Seq 2000, 180 bp PE libraries, 100 bp reads, 12 Mreads), who provided clean reads. All subsequent bioinformatics analyses were conducted in-house as previously described (Jorrin and Imperial 2014). Briefly, reads were quality filtered with Trimmomatic (Bolger et al. 2014) and were aligned against the *R. leguminosarum* bv. *viciae* 3841 reference genome (GenBank accession number PRJNA344) with Bowtie2 (Langmead and Salzberg 2012), using very sensitive standard parameters in end-to end mode (–D 20, –R 3, –N 0 –L 20 –i S, 1, 0.50). Output from the alignment was transformed with Samtools (Li et al. 2009) and SNP were detected with VarScan2.3.6 (Koboldt et al. 2012). SNP were called only if they satisfied the following conditions: minimum coverage of 20 reads, minimum frequency of 0.1, and Fisher exact test P values of observed reads vs. expected nonvariant under 0.01. The genetic flow among plant-selected subpopulations was studied by calculating pairwise Wright's fixation indices (F_{ST}) for all SNP, as calculated with the Popoolation 2 package (Kofler et al. 2011a). Those SNP with pairwise $F_{ST} > 0.1$ and $P < 0.05$ were considered to provide a significant level of differentiation between populations (Barreiro et al. 2008). For graphic representation of global SNP data, the output list of SNP for each subpopulation obtained after VarScan2 analysis was used as input to calculate Euclidean distances (dissimilarity) in SPSS (IBM, v. 20.2). The dissimilarity matrix was subjected to multidimensional scaling with the SPSS PROXCAL module and predicted coordinates were plotted in a two-dimensional space with KaleidaGraph (Synergy Software, v. 4.5.1).

Other results were graphically analyzed with Seqmonk and IGV (Robinson et al. 2011). Reference sequence coverage was expressed as number of reads or as RKPM (reads per kilobase per million reads). RKPM values were calculated as: number of reads*10^3*10^6/length*total number of reads for each subpopulation (Mortazavi et al. 2008). Percent G+C composition values along sequences were calculated with a custom script. For both RKPM coverage and percent G+C composition, sliding window sizes of 1,000 bp (Fig. 1) or 100 bp (Figs. 2 and 3) were used.

ACKNOWLEDGMENTS

This paper is dedicated to the memory of Gisèle Laguerre, who passed away in January 2013. Her work on rhizobial soil ecology inspired our research and she graciously made the P1 Dijon soil and all her research data available to us. We thank R. Ghai (Universidad Miguel Hernández, Alicante) for the script used to draw Figure 1. We also thank B. Brito and R. Prieto for their help on the initial collection of plant-selected populations and three anonymous reviewers for very insightful comments on our initial submission. This work was supported by the MICROGEN Consolider-Ingenio program (MINECO, CSD2009-00006).

LITERATURE CITED

Acosta, J. L., Eguiarte, L. E., Santamaría, R. I., Bustos, P., Vinuesa, P. Martinez-Romero, E., Davila, G., and Gonzalez, V. 2011. Genomic lineages of *Rhizobium etli* revealed by the extent of nucleotide polymorphisms and low recombination. BMC Evol. Biol. 11:305.

Aird, D., Ross, M. G., Chen, W.-S., Danielsson, M., Fennell, T., Russ, C., Jaffe, D. B., Nusbaum, C., and Gnirke, A. 2011. Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. Genome Biol. 12:R18.

Bailly, X., Giuntini, E., Sexton, M. C., Lower, R. P., Harrison, P. W. Kumar, N., and Young, J. P. W. 2011. Population genomics of *Sinorhizobium medicae* based on low-coverage sequencing of sympatric isolates. ISME J. 5:1722-1734.

Barreiro, L. B., Laval, G., Quach, H., Patin, E., and Quintana-Murci, L. 2008. Natural selection has driven population differentiation in modern

humans. Nature Genetics 40:340-345.
Bolger, A. M., Lohse, M., and Usadel, B. 2014. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics 30:2114-2120.
Denarie, J., Debelle, F., and Prome, J. C. 1996. Rhizobium lipo-chitooligosaccharide nodulation factors: Signaling molecules mediating recognition and morphogenesis. Annu. Rev. Biochem. 65:503-535.
Depret, G., Houot, S., Allard, M. R., Breuil, M. C., Nouaim, R., and Laguerre, G. 2004. Long-term effects of crop management on *Rhizobium leguminosarum* biovar viciae populations. FEMS (Fed. Eur. Microbiol. Soc.) Microbiol. Ecol. 51:87-97.
Doyle, J. J., and Luckow, M. A. 2003. The rest of the iceberg: Legume diversity and evolution in a phylogenetic context. Plant Physiol. 131:900-910.
Epstein, B., Branca, A., Mudge, J., Bharti, A. K., Briskine, R. Farmer, A. D., Sugawara, M., Young, N. D., Sadowsky, M. J., and Tiffin, P. 2012. Population genomics of the facultatively mutualistic bacteria *Sinorhizobium meliloti* and *S. medicae*. PLoS Genet. 8:e1002868. Published online.
Futschik, A., and Schloetterer, C. 2010. The next generation of molecular markers from massively parallel sequencing of pooled DNA samples. Genetics 186:207-218.
Gage, D. J. 2004. Infection and invasion of roots by symbiotic, nitrogen-fixing rhizobia during nodulation of temperate legumes. Microbiol. Mol. Biol. Rev. 68:280-300.
Graham, P. H., and Vance, C. P. 2003. Legumes: Importance and constraints to greater use. Plant Physiol. 131:872-877.
Herridge, D. F., Peoples, M. B., and Boddey, R. M. 2008. Global inputs of biological nitrogen fixation in agricultural systems. Plant Soil 311:1-18.
Hirsch, A. M. 1992. Developmental biology of legume nodulation. New Phytol. 122:211-237.
Jorrin, B., and Imperial, J. 2014. Pool-seq analysis of microsymbiont selection by the legume plant host. In: Biological Nitrogen Fixation. F. de Bruijn, ed. Wiley: Blackwell, London. In press.
Koboldt, D. C., Zhang, Q. Y., Larson, D. E., Shen, D., McLellan, M. D., Lin, L., Miller, C. A., Mardis, E. R., Ding, L., and Wilson, R. K. 2012. VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res. 22:568-576.
Koch, M., Delmotte, N., Rehrauer, H., Vorholt, J. A., Pessi, G., and Hennecke, H. 2010. Rhizobial adaptation to hosts, a new facet in the legume root-nodule symbiosis. Mol. Plant-Microbe Interact. 23:784-790.
Kofler, R., Pandey, R. V., and Schlötterer, C. 2011a. PoPoolation2: Identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). Bioinformatics 27:3435-3436.
Kofler, R., Orozco-terWengel, P., De Maio, N., Pandey, R. V., Nolte, V., Futschik, A., Kosiol, C., and Schloetterer, C. 2011b. PoPoolation: A toolbox for population genetic analysis of next generation sequencing data from pooled individuals. PLoS ONE 6:e15925. Published online.
Laguerre, G., Louvrier, P., Allard, M. R., and Amarger, N. 2003. Compatibility of rhizobial genotypes within natural populations of *Rhizobium leguminosarum* biovar viciae for nodulation of host legumes. Appl. Environ. Microbiol. 69:2276-2283.
Langmead, B., and Salzberg, S.L. 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods 9:357-359.
Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078-2079.
Long, S. R. 1996. Rhizobium symbiosis: Nod factors in perspective. Plant Cell 8:1885-1898.
Louvrier, P., Laguerre, G., and Amarger, N. 1996. Distribution of symbiotic genotypes in *Rhizobium leguminosarum* biovar viciae populations isolated directly from soils. Appl. Environ. Microbiol. 62:4202-4205.
Martinez-Romero, E. 2003. Diversity of *Rhizobium-Phaseolus vulgaris* symbiosis: Overview and perspectives. Plant Soil 252:11-23.
Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature Methods 5:621-628.
Mutch, L. A., and Young, J. P. W. 2004. Diversity and specificity of *Rhizobium leguminosarum* biovar viciae on wild and cultivated legumes. Mol. Ecol. 13:2435-2444.
Mutch, L. A., Tamimi, S. M., and Young, J. P. W. 2003. Genotypic characterisation of rhizobia nodulating *Vicia faba* from the soils of Jordan: A comparison with UK isolates. Soil Biol. Biochem. 35:709-714.
Oldroyd, G. E. D., Murray, J. D., Poole, P. S., and Downie, J. A. 2011. The rules of engagement in the legume-rhizobial symbiosis. Annu. Rev. Genet. 45:119-144.
Palmer, K. M., and Young, J. P. W. 2000. Higher diversity of *Rhizobium leguminosarum* biovar viciae populations in arable soils than in grass soils. Appl. Environ. Microbiol. 66:2445-2450.
Perret, X., Staehelin, C., and Broughton, W. J. 2000. Molecular basis of symbiotic promiscuity. Microbiol. Mol. Biol. Rev. 64:180-201.
Robinson, J. T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., and Merisov, J. P. 2011. Integrative genomics viewer. Nature Biotechnol. 29:24-26.
Sanchez-Canizares, C., Rey, L., Duran, D., Temprano, F., Sanchez-Jimenez, P., Navarro, A., Polajnar, M., Imperial, J., and Ruiz-Argueso, T. 2011. Endosymbiotic bacteria nodulating a new endemic lupine *Lupinus mariae-josephi* from alkaline soils in Eastern Spain represent a new lineage within the *Bradyrhizobium* genus. Syst. Appl. Microbiol. 34:207-215.
Schloetterer, C., Tobler, R., Kofler, R., and Nolte, V. 2014. Sequencing pools of individuals—Mining genome-wide polymorphism data without big funding. Nature Rev. Genet. 15:749-763.
Somasegaran, P. H. H. J. 1994. Handbook for Rhizobia: Methods in Legume-Rhizobium Technology. Springer-Verlag, New York.
Sugawara, M., Epstein, B., Badgley, B. D., Unno, T., Xu, L. Reese, J., Gyaneshwar, P., Denny, R., Mudge, J., Bharti, A. K., Farmer, A. D., May, G. D., Woodward, J. E., Medigue, C., Vallenet, D., Lajus, A., Rouy, Z., Martinez-Vaz, B., Tiffin, P., Young, N. D., and Sadowsky, M. J. 2013. Comparative genomics of the core and accessory genomes of 48 *Sinorhizobium* strains comprising five genospecies. Genome Biol. 14:R17.
Surin, B. P., and Downie, J. A. 1989. *Rhizobium leguminosarum* genes required for expression and transfer of host specific nodulation. Plant Mol. Biol. 12:19-29.
Vanrhijn, P., and Vanderleyden, J. 1995. The rhizobium-plant symbiosis. Microbiol. Rev. 59:124-142.
Vincent, J. M. 1970. A manual for the practical study of the root-nodule bacteria. Blackwell Scientific Publ., Oxford.
Wilson, K. 2001. Preparation of genomic DNA from bacteria. In: Current Protocols in Molecular Biology. F. M. Ausubel, ed. John Wiley & Sons, New York.
Woese, C. R. 1987. Bacterial evolution. Microbiol. Rev. 51:221-271.
Woese, C. R., Kandler, O., and Wheelis, M. L. 1990. Towards a natural system of organisms - proposal for the domains archaea, bacteria, and eucarya. Proc. Natl. Acad. Sci. U.S.A. 87:4576-4579.
Young, J. P. W., Crossman, L. C., Johnston, A. W. B., Thomson, N. R., Ghazoui, Z. F., Hull, K. H., Wexler, M., Curson, A. R. J., Todd, J. D., Poole, P. S., Mauchline, T. H., East, A. K., Quail, M. A., Churcher, C., Arrowsmith, C., Cherevach, I., Chillingworth, T., Clarke, K., Cronin, A., Davis, P., Fraser, A., Hance, Z., Hauser, H., Jagels, K., Moule, S., Mungall, K., Norbertczak, H., Rabbinowitsch, E., Sanders, M., Simmonds, M., Whitehead, S., and Parkhill, J. 2006. The genome of *Rhizobium leguminosarum* has recognizable core and accessory components. Genome Biol. 7:R34.